

Impossible & Advanced Optical Character Recognition Using a Cheap Point & Shoot Digital Camera

INTRODUCTION

This study was conducted using a point & shoot digital camera *Konica Minolta DiMAGE XI* with 8.1 Megapixels with no previous adjustment at 3264x2448 pixels with 72 dpi. Samples from 4 books were used:

1. [MILTON, JOHN.] *The Grand Case of Conscience concerning the Engagement Stated & Resolved.*
printed in London 1650
2. [PETKOV-MISIRKOV, KRSTE.] *Za makedonskite raboti.*,
printed in Sofia 1903
3. [DE MONTESQUIEU, MONSIEUR.] *Oeuvres. Tome I,*
printed in London 1787
4. [VARIOUS AUTHORS] *Macedonian Review, year IV, book#4,*
printed in Sofia 1933

Only JONH MILTON book was "scanned" at Alexander Library at Rutgers, the State University of New Jersey with expensive professional camera scanner. It is used here for the sake of the comparison in relation to the curve of the pages and the **high resolution professional cameras Vs cheap point & shoot digital cameras.**

The "scanning" was performed using a light bulb of 50W in room conditions. Some pattern training was done using Abbyy Finereader 6.0 Different compression methods were used to reduce the file size to its minimum web centricity. Optical Character Recognition engines require the minimum of 150 dpi (the optimum is 300 dpi) for accurate results. Using an experimental program with Abbyy Finereader 5.0 engine at 72 dpi JPEG (instead of TIF) files some amazing results appeared!

RESULTS

When making a full text search of the files by using the actual alphabet, the recognized text is given/written in odd or non-existing characters. Having in mind that very cheap equipment was used in everyday (ordinary) conditions, the quality of the extracted text is excellent. There is no explanation needed in this case (please look at the screenshots on the poster very carefully.)

CONCLUSION

It is possible to extract the text by using a cheap point and shoot digital camera with 8 MP and by some previous experimentation of the light and the curvature of the pages. If you consult an Optical Character Recognition Specialist, it will save you a lot of trouble and time when you scan or photograph rare and old books. Most of the libraries, national archives and museums use digital cameras instead of scanners to "scan" the books. It is very complicated to extract the text out of them, because of the 72dpi resolution and the "no adjustments features" made by the person who uses the expensive equipment for photographing the book.

FUTURE

There is more to be done in using the digital camera to digitalize a book. The high resolution camera scanner does not guarantee the best results. We need more research in the field of the scanning and the right resolution, curve of the pages, ISO settings, exposure time and bigger lenses.

Can 72 dpi (half than the minimum for OCR) of the point & shoot be better than expensive and sometimes slow scanners? **Yes!** Do we need special conditions to increase the quality of the OCR? **No, we don't!** Library users, in a year or two, will be able to digitize a book at their library work desk using a cheap digital camera, free software or free internet conversion servers to DJVu format or demo version of the ABBYY Finereader software (ScanR servers uses only 2 MP pictures from the mobile phone for OCR).

Also, please remember: OCR done with the digital camera is at 72dpi...The minimum is 150 dpi,
optimum 300dpi. The future of OCR is right here-right now...Thank you!

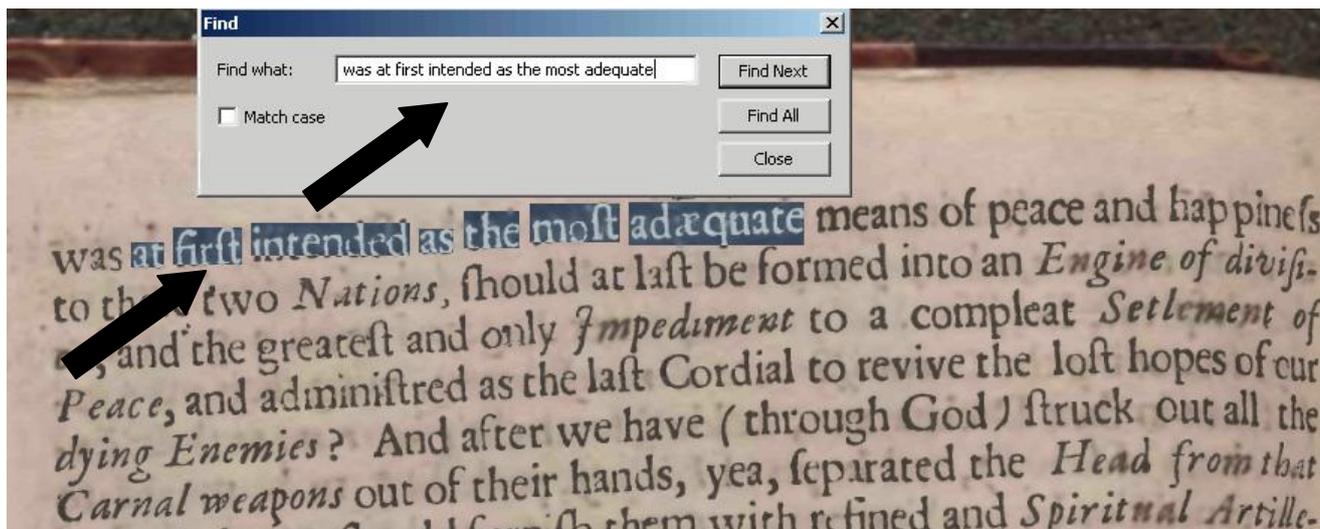
Poposki Dimitar

OCR Specialist for Rare Books and Special Collections

US Department of State Alumni Junior Faculty Development Program

Vasil Djorgov 28/16 1000 Skopje, Macedonia

Work: +389 2 3240-492 Fax: +389 2 3223-811 Home: +389 2 3230-604 Mobile: +389 75 855540



Sample from [MILTON, JOHN.] *The Grand Case of Conscience concerning the Engagement Stated & Resolved*, printed in London 1650.

Professional scanner camera is used. Some of the characters are not well recognized by the OCR machine. Still, the book can be searched in modern English characters.

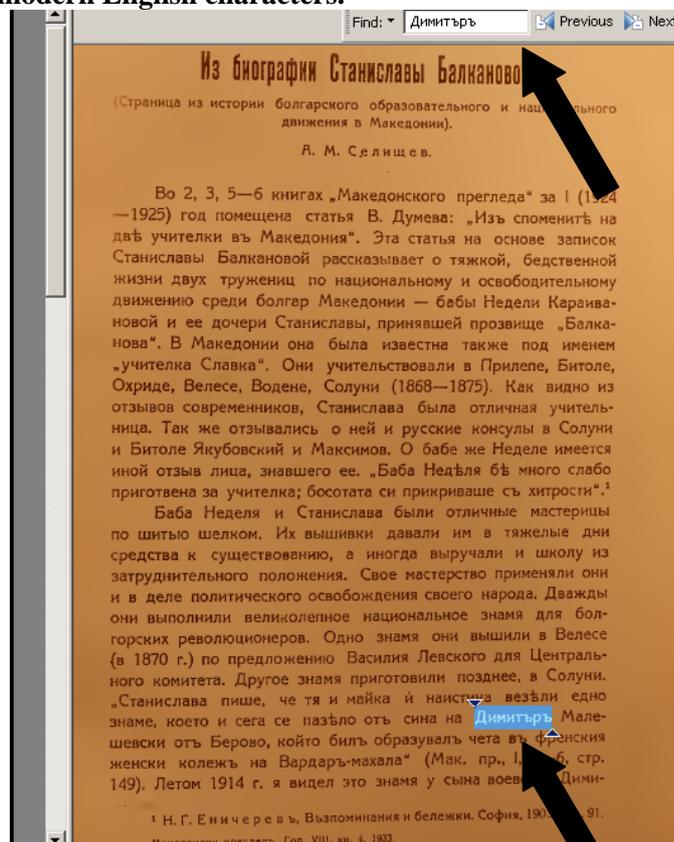
Screen 1 of 2

Из биографии Станиславы Балкановой.

(Страница из истории болгарского образовательного и национального движения в Македонии).

Я. М. Селищев.

Во 2, 3, 5—6 книгах „Македонского прегледа“ за 1 (1924—1925) год помещена статья В. Думева: „Изъ спомениѣ на дваѣ учителки въ Македония“. Эта статья на основе записок Станиславы Балкановой рассказывает о тяжелой, бедственной жизни двух тружениц по национальному и освободительному движению среди болгар Македонии — бабы Недели Караивановой и ее дочери Станиславы, принявшей прозвище „Балканова“. В Македонии она была известна также под именем „учителка Славка“. Они учительствовали в Прилепе, Битоле, Охриде, Велесе, Водене, Солуни (1868—1875). Как видно из отзывов современников, Станислава была отличная учительница. Так же отзывались о ней и русские консулы в Солуни и Битоле Якубовский и Максимов. О бабе же Неделе имеется иной отзыв лица, знавшего ее. „Баба Неделя бѣ много слабо пригтовена за учителка; босотата си прикриваше съ хитрости“. 1 Баба Неделя и Станислава были отличные мастерицы по шитью шелком. Их вышивки давали им в тяжелые дни средства к существованию, а иногда выручали и школу из затруднительного положения. Свое мастерство применяли они и в деле политического освобождения своего народа. Дважды они выполнили великолепное национальное знамя для болгарских революционеров. Одно знамя они вышли в Велесе (в 1870 г.) по предложению Василия Левского для Центрального комитета. Другое знамя приготовили позднее, в Солуни. „Станислава пише, че тя и майка и наистина везли едно знаме, което и сега се пазило отъ сына на Димитър Малешевски отъ Берово, който билъ образувалъ чета въ френския женски колежъ на Вардаръ-махала“ (Мак. пр., I, 5—6, стр. 149). Летом 1914 г. я видел это знамя у сына воеводы Димитрия Н. Г. Еничерева, Възпоминания и белешки. София, 1906, стр. 91. Македонски прегледъ. Год. VIII, кн. 4, 1933.



Sample from [VARIOUS AUTHORS] *Macedonian Review*, year IV, book#4, printed in Sofia 1933. Made with point & shoot digital camera *Konica Minolta DiMAGE X1* with 8.1 Megapixels with no previous adjustment at 3264x2448 pixels with 72 dpi.

For the sake of the comparison on the left is extracted text and on the right the full text search original.

Poposki Dimitar

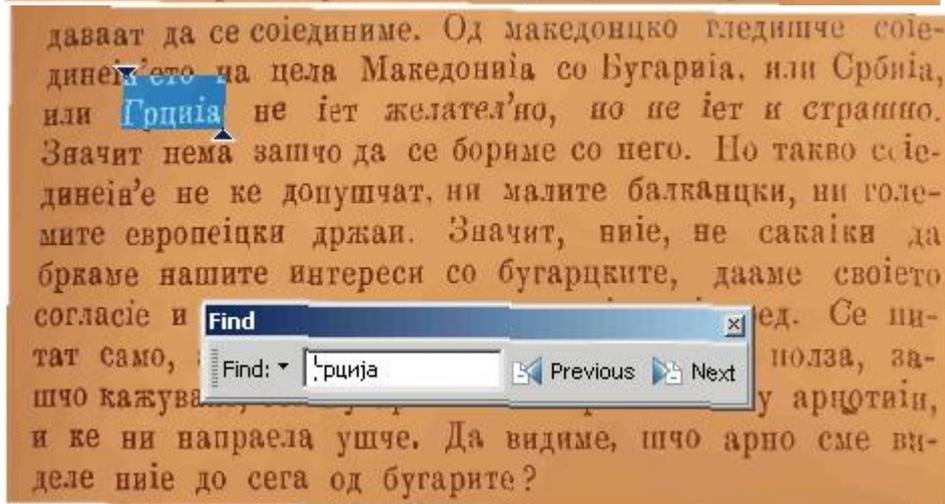
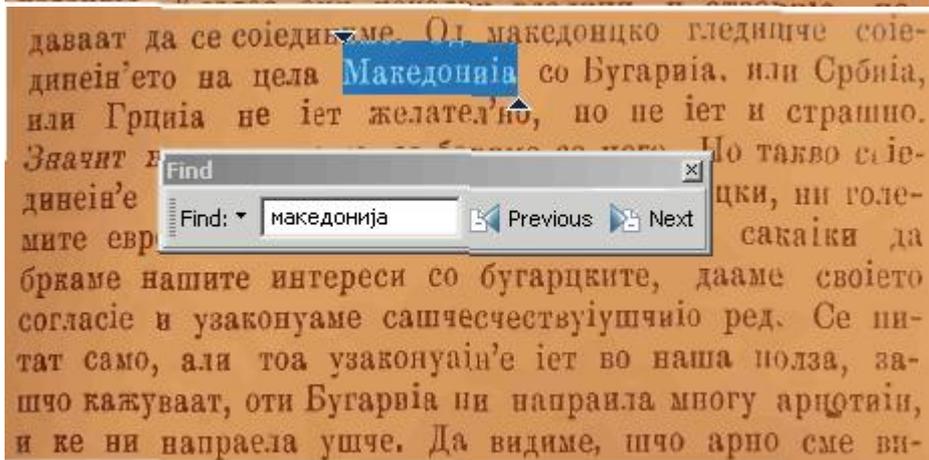
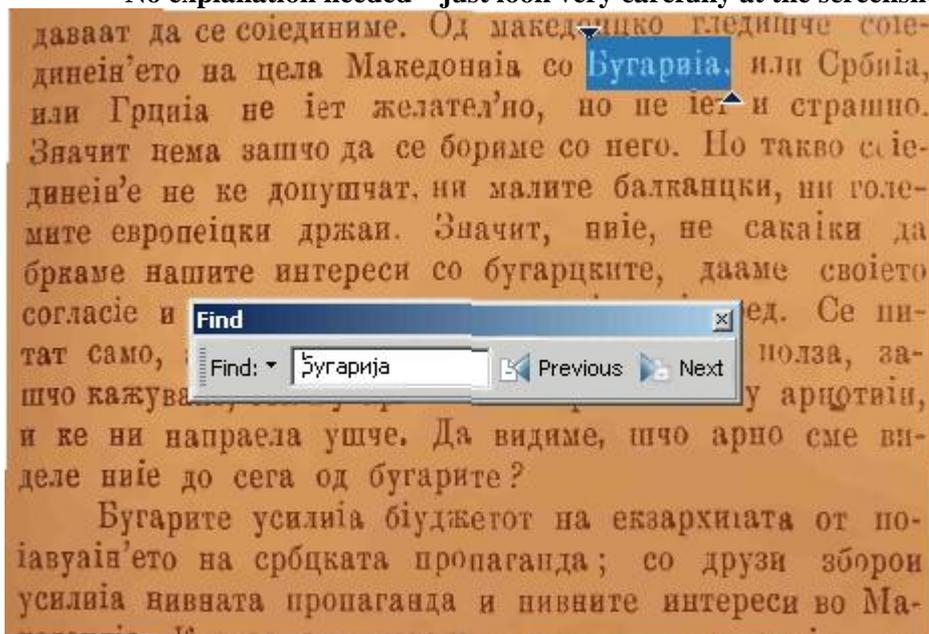
OCR Specialist for Rare Books and Special Collections

US Department of State Alumni Junior Faculty Development Program

Vasil Djorgov 28/16 1000 Skopje, Macedonia

Work: +389 2 3240-492 Fax: +389 2 3223-811 Home: +389 2 3230-604 Mobile: +389 75 855540

Advanced Optical Character recognition done with point and shoot digital camera at 72dpi.
No explanation needed – just look very carefully at the screenshots samples.



[PETKOV-MISIRKOV, KRSTE.] Za makedonskite raboti., printed in Sofia 1903

Poposki Dimitar

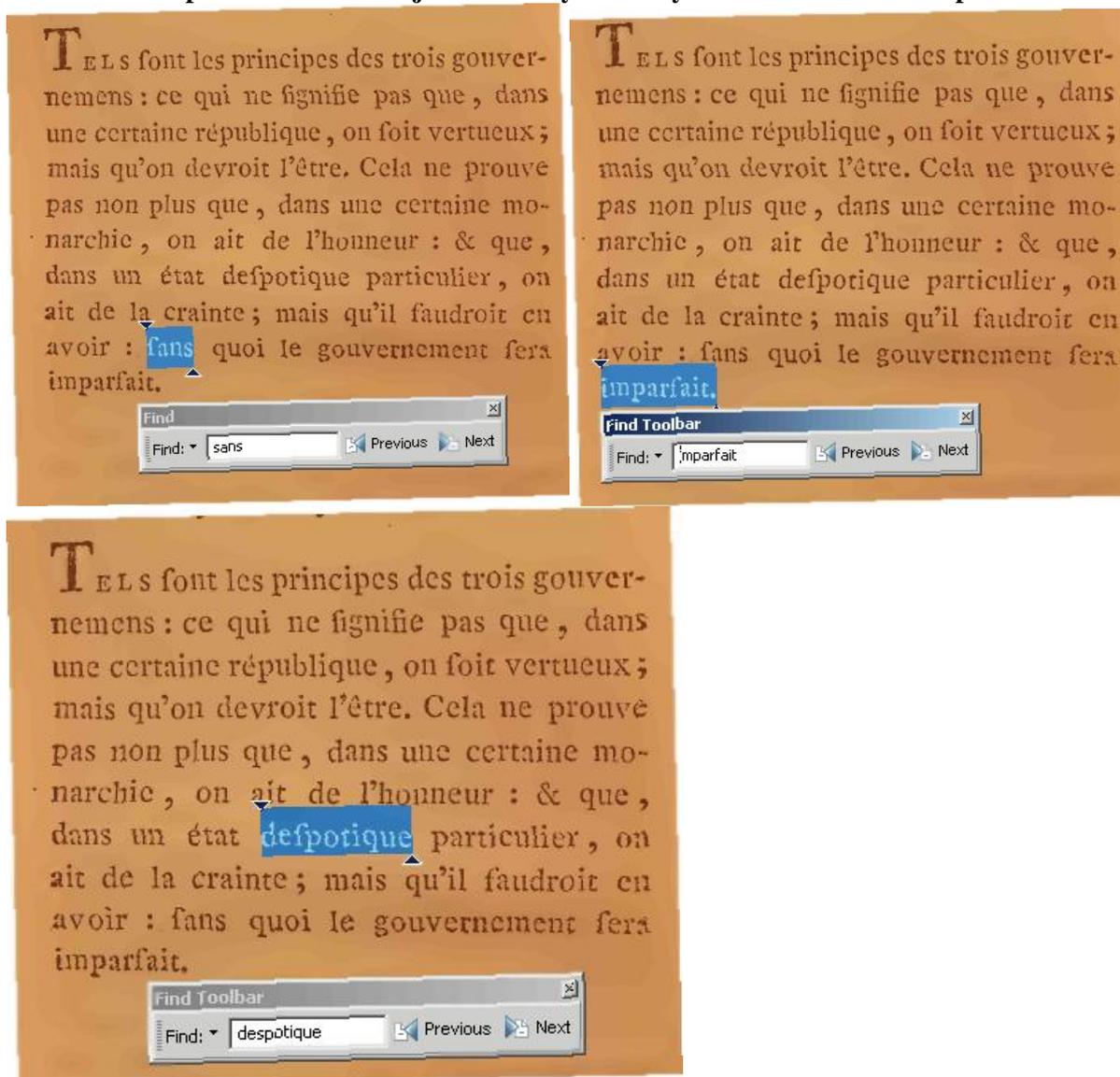
OCR Specialist for Rare Books and Special Collections

US Department of State Alumni Junior Faculty Development Program

Vasil Djorgov 28/16 1000 Skopje, Macedonia

Work: +389 2 3240-492 Fax: +389 2 3223-811 Home: +389 2 3230-604 Mobile: +389 75 855540

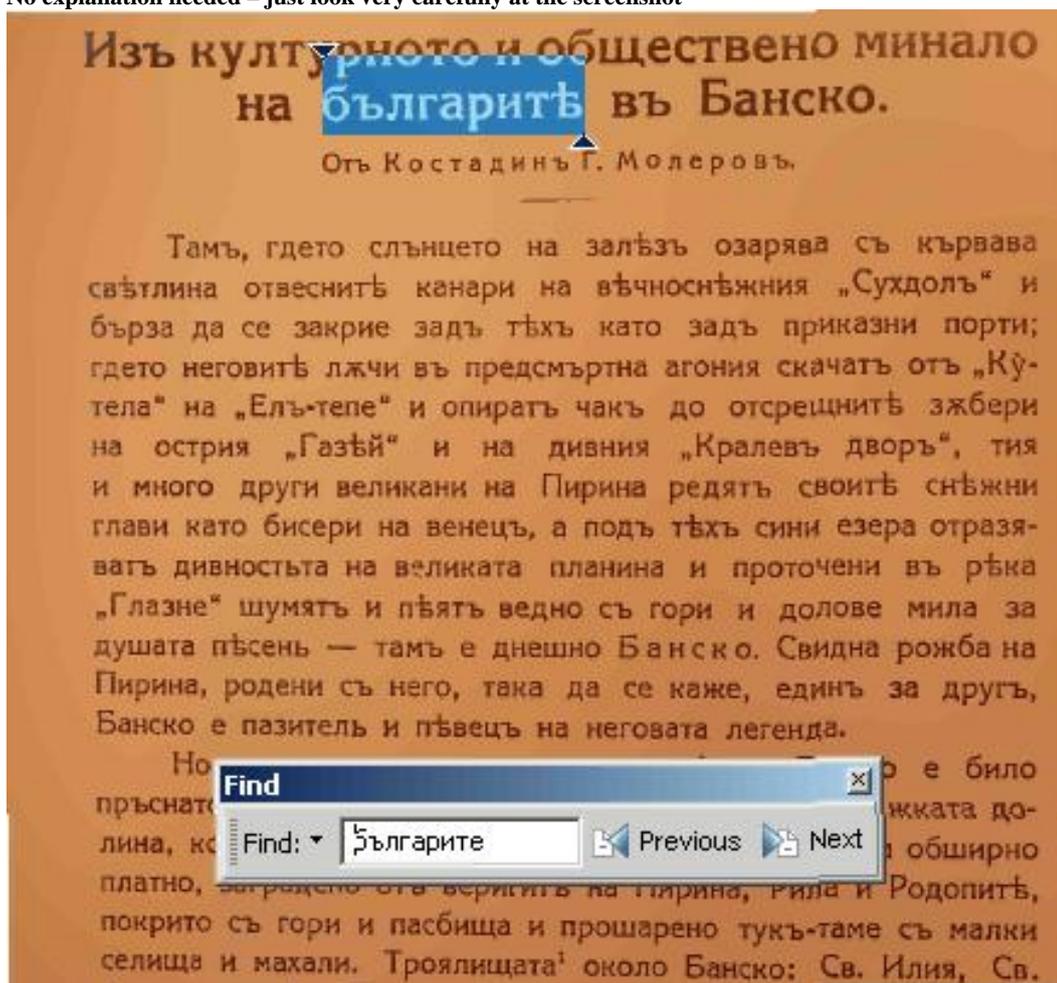
**Advanced Optical Character recognition done with point and shoot digital camera at 72dpi.
No explanation needed – just look very carefully at the screenshots samples.**



[DE MONTESQUIEU, MONSIEUR.] Oeuvres. Tome I, printed in London 1787

Poposki Dimitar
OCR Specialist for Rare Books and Special Collections
 US Department of State Alumni Junior Faculty Development Program
 Vasil Djorgov 28/16 1000 Skopje, Macedonia
 Work: +389 2 3240-492 Fax: +389 2 3223-811 Home: +389 2 3230-604 Mobile: +389 75 855540

Advanced Optical Character recognition done with point and shoot digital camera at 72dpi.
No explanation needed – just look very carefully at the screenshot



Изъ културното и обществено минало на българитѣ въ Банско.

Отъ Костадинъ Г. Молеровъ.

Тамъ, гдето слънцето на залѣзъ озарява съ кървава свѣтлина отвеснитѣ канари на вѣчноснѣжния „Сухдолъ“ и бърза да се закрие задъ тѣхъ като задъ приказни порти; гдето неговитѣ лжчи въ предсмъртна агония скачатъ отъ „Кутела“ на „Елъ-тепе“ и опиратъ чакъ до отсрещнитѣ зжбери на острия „Газѣй“ и на дивния „Кралевъ дворъ“, тия и много други великани на Пирина редятъ своитѣ снѣжни глави като бисери на венецъ, а подъ тѣхъ сини езера отразяватъ дивността на великата планина и проточени въ рѣка „Глазне“ шумятъ и пѣятъ ведно съ гори и долове мила за душата пѣсенъ — тамъ е днешно Банско. Свидна рожба на Пирина, родени съ него, така да се каже, единъ за другъ, Банско е пазителъ и пѣвецъ на неговата легенда.

Но преди да заеме днешното си мѣсто, Банско е било пръснато гката долина, която обширно платно, задопитѣ, покрито съ гори и пасоища и прошарено тукъ-таме съ малки



Изъ културното и обществено минало на българитѣ въ Банско.

Отъ Костадинъ Г. Молеровъ.

Тамъ, гдето слънцето на залѣзъ озарява съ кървава свѣтлина отвеснитѣ канари на вѣчноснѣжния „Сухдолъ“ и бърза да се закрие задъ тѣхъ като задъ приказни порти; гдето неговитѣ лжчи въ предсмъртна агония скачатъ отъ „Кутела“ на „Елъ-тепе“ и опиратъ чакъ до отсрещнитѣ зжбери на острия „Газѣй“ и на дивния „Кралевъ дворъ“, тия и много други великани на Пирина редятъ своитѣ снѣжни глави като бисери на венецъ, а подъ тѣхъ сини езера отразяватъ дивността на великата планина и проточени въ рѣка „Глазне“ шумятъ и пѣятъ ведно съ гори и долове мила за душата пѣсенъ — тамъ е днешно Банско. Свидна рожба на Пирина, родени съ него, така да се каже, единъ за другъ, Банско е пазителъ и пѣвецъ на неговата легенда.

Но преди да заеме днешното си мѣсто, Банско е било пръснато гката долина, която обширно платно, задопитѣ, покрито съ гори и пасоища и прошарено тукъ-таме съ малки селища и махали. Троялицата¹ около Банско: Св. Илия, Св.



[VARIOUS AUTHORS] Macedonian Review, year IV, book#4, printed in Sofia 1933

Poposki Dimitar

OCR Specialist for Rare Books and Special Collections

US Department of State Alumni Junior Faculty Development Program

Vasil Djorgov 28/16 1000 Skopje, Macedonia

Work: +389 2 3240-492 Fax: +389 2 3223-811 Home: +389 2 3230-604 Mobile: +389 75 855540